c246 Problem Set 7

Deadline: due 15 March 2002, 5pm by email to homework@c246.lbl.gov

Notes:
1. This is really the second half of problem set 6, but because it was getting to long, we have broken it into two parts and provided additional time to complete the problem set. Each problem set is worth as much as a normal full problem set.
2. Turning in problem set early is encouraged so you can get feedback well in advance of the exam.
3. Answer key will be posted shortly after homework is due, to allow review before the exam. Thus, late problem sets will receive no credit.

1) (10 points) What are sources of errors in functional annotation of protein sequences? What is their impact? How can you recognize errors in the annotation of proteins?

2) (15 points) Why would you filter your query sequence using a program like ccp / coils when doing a database search? What other programs might you also use for comparable purposes?

3) (15 points) Given two sets of 5 sequences of length 500, sometimes MSA will run slowly and sometimes it will run very slowly. What features of the sequences and their alignments would cause this to happen?

4) (10 points) Why are hand-edited multiple alignments often superior to automated alignments?

5) (10 points) What is the meaning of "once a gap, always a gap"? Name 3 programs that have this "feature" and 5 that do not.

6) (10 points) What are the key reasons that rigorous traditional sequence alignment by dynamic programming becomes intractable with multiple sequences? What are advantages of the full DP over the other approaches?

7) (30 points) From the alignment you made in Problem Set 6, Part I, extract the alignment from positions 51-55. Build a profile / PSSM for this multiple alignment using (a) the average method, and (b) the data-dependent pseudocount approach used by PSI-BLAST. Either write a program to do this or show your work.